Statistics for Biology and Health

Borchers/Buckland/Zucchini: Estimating Animal Abundance: Closed Populations.
Burzykowski/Molenberghs/Buyse: The Evaluation of Surrogate Endpoints.
Everitt/Rabe-Hesketh: Analyzing Medical Data Using S-PLUS.
Ewens/Grant: Statistical Methods in Bioinformatics: An Introduction. 2nd ed.
Hougaard: Analysis of Multivariate Survival Data.
Keyfitz/Caswell: Applied Mathematical Demography, 3rd ed.
Klein/Moeschberger: Survival Analysis: Techniques for Censored and Truncated Data, 2nd ed.
Kleinbaum/Klein: Survival Analysis: A Self-Learning Text.
Kleinbaum/Klein: Survival Analysis: A Self-Learning Text. 2nd ed.
Kleinbaum/Klein: Logistic Regression: A Self-Learning Text, 2nd ed.
Lange: Mathematical and Statistical Methods for Genetic Analysis, 2nd ed.
Manton/Singer/Suzman: Forecasting the Health of Elderly Populations.
Moyé: Multiple Analyses in Clinical Trials: Fundamentals for Investigators.
Parmigiani/Garrett/Irizarry/Zeger: The Analysis of Gene Expression Data: Methods and Software.
Salsburg: The Use of Restricted Significance Tests in Clinical Trials.
Simon/Korn/McShane/Radmacher/Wright/Zhao: Design and Analysis of DNA Microarray Investigations.
Sorensen/Gianola: Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics.
Stallard/Manton/Cohen: Forecasting Product Liability Claims: Epidemiology and Modeling in the Manville Asbestos Case.
Therneau/Grambsch: Modeling Survival Data: Extending the Cox Model.
Vittinghoff/Glidden/Shiboski/McCulloch: Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models
Zhang/Singer: Recursive Partitioning in the Health Sciences.

Tomasz Burzykowski
Geert Molenberghs
Marc Buyse
Editors

The Evaluation of Surrogate Endpoints

With 57 Illustrations

Springer

20

The Promise and Peril of Surrogate Endpoints in Cancer Research

Arthur Schatzkin, Mitch Gail, and Laurence Freedman

20.1 Introduction

Cancer is one of humanity's leading causes of morbidity and mortality. Nevertheless, in the general population, even the most common malignancies have a low probability of occurrence over a restricted time interval. For example, the age-adjusted annual incidence rate of breast cancer among women in the United States is about 100 per 100,000, or 0.1%; the annual colorectal cancer incidence rate among men and women combined is around 50 per 100,000, or only 0.05%. And these are among the most frequently occurring malignancies.

The medical research implications of this relative infrequency of cancer occurrence are straightforward: controlled intervention studies or prospective observational epidemiologic investigations that use incident cancer as an endpoint must be large, lengthy, and, therefore, costly. Such studies must yield many hundreds of cancers to have adequate statistical power to detect a meaningful treatment effect or exposure association. The ongoing Women's Health Initiative, for example, requires several tens of thousands of participants to be followed over nearly a decade to observe sufficient numbers of cancers to detect reasonable reductions in the incidence of breast and colorectal malignancies (Women's Health Initiative Study Group 1998). Studies with surrogate endpoints, biomarkers of preclinical carcinogenesis, are attractive because such studies are potentially smaller, shorter, and considerably less expensive than their counterparts with cancer endpoints.

20.2 When Are Surrogates Appropriate?

Despite their potential to reduce the size, duration, and cost of studies, surrogate endpoints may not be acceptable because the quality of evidence they provide on treatment effects or exposure associations is lower than that obtained by studying the effects of treatment or exposure on a true cancer endpoint. For some types of studies, the quality of evidence provided by surrogates might be sufficient, whereas for others only the cancer endpoints will do. For example, true clinical endpoints, such as time to cancer recurrence or time to death, might be indispensable in randomized phase III clinical trials designed to estimate the clinical effects of a new cancer treatment. Such trials must provide the highest standards of evidence regarding treatment efficacy. Phase II trials, on the other hand, are preliminary studies designed to determine whether an agent warrants further study in phase III trials, so the use of a surrogate endpoint, such as whether a tumor shrinks following treatment, might be acceptable. The consequences of a false negative result might be to curtail testing of a potentially valuable treatment; a false positive result would not lead to widespread use of the agent, however, but only to phase III testing, where, presumably, the agent would be found to have no beneficial clinical effect. Likewise, in epidemiologic investigations of, for example, the relationship of dietary factors to colorectal or breast cancer, surrogate endpoints such as cell proliferation indices or blood hormone concentrations might provide valuable exploratory information in the evaluation of a new hypothesis, whereas more rigorous testing of that dietary hypothesis might require the use of frank cancer endpoints.

20.3 Identifying Surrogate Endpoints for Cancer

To define a surrogate endpoint (S), it is necessary first to define the true clinical endpoint (T). In most observational epidemiologic studies, T is the occurrence of new ("incident") cancer, usually specified as the age or time of cancer diagnosis. In therapeutic clinical trials, T is usually taken as the time from treatment to either cancer recurrence or death. Other clinically meaningful measures that influence how a patient feels or functions can also be used as primary endpoints (DeGruttola *et al.* 2004). Any measurement other than T is a potential surrogate measurement. In a preamble to a proposed accelerated approval rule for drugs, the Food and Drug Administration defined a surrogate as follows: "A surrogate endpoint, or 'marker', is a laboratory measurement or physical sign that is used in therapeutic trials as a substitute for a clinically meaningful endpoint that is a direct measure of how a patient feels, functions, or survives and is expected to predict the effect of the therapy" (Federal Register 1992).

There are a host of biological phenomena, potential biomarkers of preclinical carcinogenesis, that could potentially serve as cancer surrogates. With the explosion in molecular and cell biology, this list is growing:

Alterations in the characteristics of tissues. "Pre-neoplastic" or frankly neoplastic changes are obvious candidates for surrogate endpoints. Examples include cervical (Mitchell *et al.* 1994), prostatic (Bostwick 1999), and endometrial (Mutter 2000) intraepithelial neoplasia; colorectal adenomatous polyps (Schatzkin *et al.* 1994); bronchial metaplasia (a possible pre-neoplastic state for lung cancer) (Misset *et al.* 1986); and dysplastic changes in the esophagus (Dawsey *et al.* 1998).

Histological changes detected by imaging. Examples include mammographic parenchymal patterns as a surrogate for breast carcinogenesis (Saftlas *et al.* 1989), and ovarian ultrasound abnormalities in ovarian cancer (Karlan 1995).

Cellular phenomena. Surrogates in this category include several assays of epithelial cell proliferation, including tritiated thymidine or bromodeoxyuridine incorporation into DNA, proliferating cell nuclear antigen (PCNA), and Ki67 (Baron *et al.* 1995b). Measures of apoptosis (Bedi *et al.* 1995) have recently been proposed as potential surrogate endpoints, as well as the ratio of proliferation to apoptosis. In AIDS research, CD4 cell counts and HIV viral load have been used as surrogates for critical AIDS endpoints (Tsiatis, DeGruttola, and Wulfsohn 1995, Ruiz *et al.* 1996).

Molecular markers. A plethora of potential molecular surrogates have been suggested. Examples include specific somatic mutations in cancer-related genes (such as RAS or TP53), DNA *hypo-* and *hyper*-methylation of specific genes, and gene expression products (including those measured in microarrays) (Fearon 1992, Counts and Goodman 1995, Brown and Botstein 1999). Chemical-DNA adducts can be considered not only as indicators of exposure (which they might well be) but also as markers of a "downstream" integrated metabolic process, one occurring temporally and developmentally closer to the malignant outcome than the exposure itself (Groopman *et al.* 1994).

Infection and inflammation. Infectious processes have been implicated in a number of cancers, and these infections could be viewed as surrogate endpoints. Examples include infections with human papillomavirus (HPV) in cervical carcinogenesis (Schiffman 1992), *Helicobacter pylori* in gastric cancer (Muñoz 1994), and HTLV1 in adult T-cell leukemia (Blattner 1989). Inflammatory cells and cytokines, which contribute to tumor growth, progression, and immunosuppression, could serve as surrogate markers (Balkwill and Mantovani 2001).

Bioactive substances in blood and tissue. Examples here include blood and tissue estrogens or androgens, oxidation products, and anti-oxidants (again, in both blood and specific tissues), tissue- or cell-type-specific antigens (such as prostate-specific antigen, PSA), and growth factors. For this category of potential surrogates, the marker, blood estrogen levels (Dorgan *et al.* 1996), for example, may not be found directly in the target tissue, but may still properly be considered a potential surrogate endpoint, in this case, for breast cancer.

Cancer prognostic factors. Potential surrogate endpoints in cancer treatment studies include time to cancer recurrence (when the true endpoint is survival) and initial tumor shrinkage (instead of true endpoint like time to tumor recurrence or survival).

20.4 Validating Surrogate Markers

Once we have found a potential surrogate, how do we determine whether it is a good surrogate marker for the true endpoint? A potential use of the surrogate, S, in assessing the effect of the exposure or intervention, E, on T is through a hypothesis test of an association between S and E. For S to be valid for hypothesis testing, the condition "S is not associated with E" (the "null hypothesis") must imply that "T is not associated with E," and vice versa (Prentice 1989). Later we discuss three conditions that are required to establish this criterion: first, S must influence T; second, E must influence S; and third, S "mediates" the effect of E on T (that is, in statistical terms, T is unrelated to E conditional on S). If S is valid for hypothesis testing, we know that if we reject the null hypothesis that S is associated with E (i.e., we accept that S is associated with E), we can conclude that T is also probably associated with E.

Although validity of hypothesis testing based on S is desirable, it would be even more useful if we could predict the magnitude of the effect of E on T from data on the magnitude of the effect of E on S. Recent proposals for such prediction are based on analyzing a series of studies of treatments in a similar class of treatments (Daniels and Hughes 1997, Buyse *et al.* 2000a, Gail *et al.* 2000), and "trial-level validity" gives an indication of how reliably one can predict the magnitude of the effect of E on T.

Suppose in each study we have sufficient information to allow us to estimate the effect of an exposure, E on a surrogate endpoint, S and the effect of E on the frank endpoint, T. We might call these two estimated treatment effects or exposure associations $\widehat{\beta}_S$ and $\widehat{\beta}_T$ obtained by regressing S on E and T on E, respectively. In Figure 20.1, pairs $(\widehat{\beta}_S, \widehat{\beta}_T)$ are plotted for seven different hypothetical clinical trials of various cancer treatments focused on the same molecular pathway, each compared with placebo. If the squared correlation, R^2, among these trial-level pairs was high, we would conclude that the effects of E on S are highly predictive of the effects of E on T, and we would say that S is "trial-level valid" (Bostwick 1999, Mutter 2000) if R^2 was near 1.0. An analysis of such a series of studies with high R^2 gives us some empirical evidence that if we wish to study a new agent in this same class of agents, we can combine data on the effect of the new agent E on S with the data from previous studies, as represented in the figure, to predict what the effect of E is on T. There are, however, a number of limitations to relying on this strategy (Schatzkin *et al.* 1994), including potentially serious loss of precision in estimates of the effect of E on T for the new agent and uncertainty about whether the new agent really belongs to the same class of agents depicted in Figure 20.1.

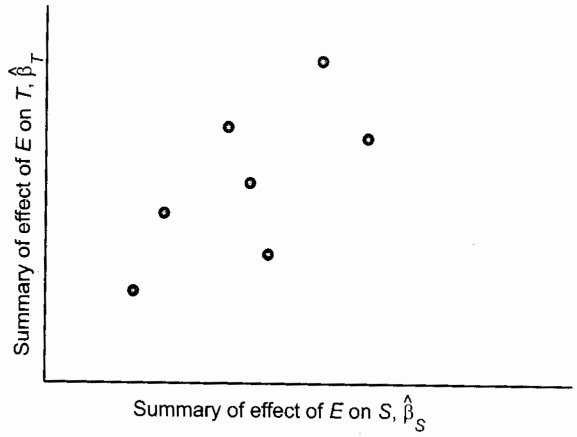

FIGURE 20.1. *Pairs of treatment effects for seven different hypothetical trials.*

We now turn to some examples that give insight into these criteria for validating a surrogate marker.

20.5 The Logic of Cancer Surrogacy

Suppose, in Figure 20.2a, $E1$ represents an "exposure" to some environmental or host factor, anything from a chemopreventive agent to a deleterious

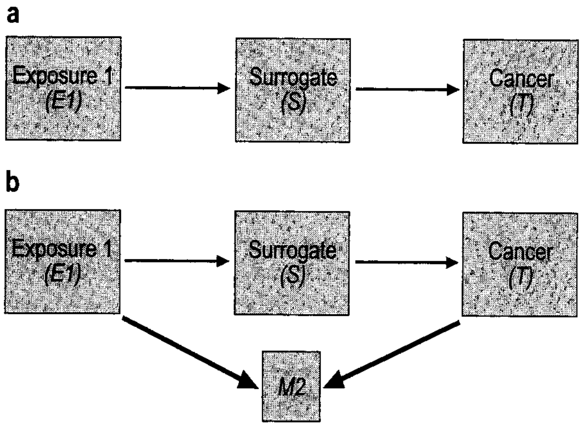

FIGURE 20.2. *Hypothetical.*

risk factor. According to this idealized model, a change in $E1$ necessarily alters the surrogate endpoint (S), which in turn modifies the true endpoint, the likelihood of incident cancer (T). As we discuss in the next section, a causal pathway such as that depicted in Figure 20.2a implies that S is valid for hypothesis testing for the particular factor $E1$, but, without further assumptions, does not necessarily imply that S will be valid for hypothesis tests for another exposure, $E2$, nor that the magnitudes of the effects of $E1$ on S can reliably predict the magnitudes of the effects of $E1$ on T for a series of exposures (trial-level validity, as described in the previous section).

The scenario in Figure 20.2a rarely occurs. Far more realistic are situations reflected in Figure 20.2b. Here, $E1$ modulates carcinogenesis through two alternative pathways, one through S, the other through another marker $M2$. To the extent that $E1$ operates through the alternative $M2$ pathway, which means that S is not a necessary component of carcinogenesis, we cannot be assured that S is a valid surrogate for hypothesis testing in studies of $E1$. The reason for this lack of certainty is that $E1$ might influence $M2$ in a way that offsets its effect on S, the final effect on cancer simply being unknown. If $E1$, for example, were to increase $M2$-positivity, $E1$ could actually end up *increasing* cancer incidence, while at the same time reducing S-positivity and giving at least a superficial impression of being anti-carcinogenic. An example from cardiovascular disease is instructive. High-dose diuretics lower blood pressure but have little effect on cardiovascular disease mortality in hypertensive patients, possibly because diuretics cause hypokalemia, which increases risk of sudden death (Temple 1999). The relationships in Figure 20.2b also make trial-level validity less likely than in Figure 20.2a, because the magnitude of the effects of E on T are less likely to be predictable from the effects of E on S in a series of such studies.

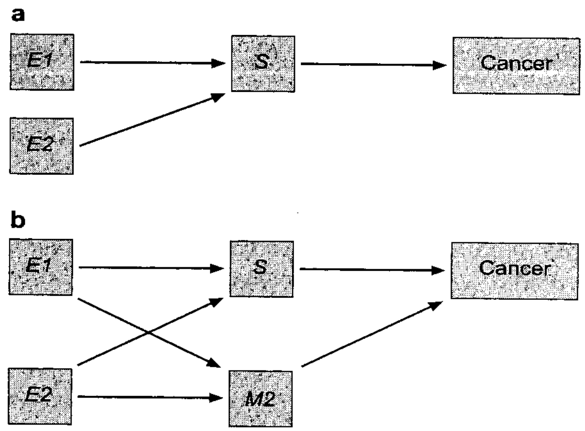

FIGURE 20.3. *Hypothetical setting.*

20.6 Can Surrogate Validity Be Extrapolated from One Exposure to Another?

Another important question is whether a surrogate that is valid for one intervention (or exposure) is valid for another. Figure 20.3a reprises Figure 20.2a but adds another exposure, $E2$. Exposure here can refer to an intervention agent or a risk factor. Both $E1$ and $E2$ operate through a single surrogate on the path to cancer. In this scenario, the surrogate is a necessary component of the cancer pathway. $E2$ must operate through the surrogate. The surrogate is valid for studies of $E2$ as well as those of $E1$.

In Figure 20.3b, $E2$ enters into the more complex scenario depicted in Figure 20.2b. The existence of a non-trivial alternative pathway (through $M2$) means that the validity of the surrogate S may be exposure dependent. Even if $E1$ works primarily through the surrogate and affects $M2$ minimally, suggesting that the surrogate is reasonably valid for $E1$-cancer studies, it cannot be assumed that the $E2 - M2$ cancer pathway plays a similarly minor role in carcinogenesis.

For example, a given agent, $E1$, might influence colorectal carcinogenesis largely through its influence on cell proliferation. Cell proliferation in this scenario is a likely valid surrogate for colorectal cancer. A second agent,

$E2$, might have a minimal effect on cell proliferation but could increase apoptosis sufficiently to decrease cancer incidence. Focusing only on cell proliferation would give a falsely pessimistic impression of the efficacy of the second agent. The validity of a surrogate must therefore be established for every intervention.

An approach to this problem is to consider studies of a "class" of biologically comparable intervention agents. If, for example, a meta-analysis shows that the effect of these agents on the surrogate predicts their effect on the true endpoint, we can be reasonably confident in inferring a treatment effect on the true endpoint from the effect of a new member of that class on the surrogate endpoint, as discussed above (Bostwick 1999, Mutter 2000, Schatzkin *et al.* 1994).

20.7 Epithelial Hyperproliferation: A Case Study

How can we apply this logic to potential surrogates? Cell proliferation assays have been touted as potential surrogates for cancer in light of the dysregulation of cell growth that characterizes malignancy (Wargovich 1996). But are they valid surrogates? Figure 20.4 depicts causal events potentially involved in the relationship between hyperproliferation and the neoplastic process in the colorectum. If we focus just on the upper portion of this diagram, we see a single pathway going from normal epithelium to hyperproliferative epithelium to neoplasia/cancer. It is this pathway that implicitly underlies using hyperproliferation as a surrogate for cancer in testing whether there is an association between an exposure and cancer.

But hyperproliferation may not be necessary by itself for colorectal carcinogenesis. There may be an alternative pathway to neoplasia/cancer that bypasses hyperproliferation. The problem is that the effect of an intervention agent ($E1$) on this alternative pathway is unknown and may in fact counterbalance the effect through the hyperproliferation pathway. Two scenarios here are revealing:

1. The agent ($E1$) reduces proliferation, but at the same time reduces apoptosis, and therefore has no effect on colorectal cancer;
2. The agent has no effect on proliferation but does increase apoptosis, thereby reducing colorectal cancer incidence.

In both cases, a hyperproliferation assay gives the wrong answer about an intervention's effect on colorectal cancer; by definition, hyperproliferation would not be a valid surrogate for testing for an association between $E1$ and cancer.

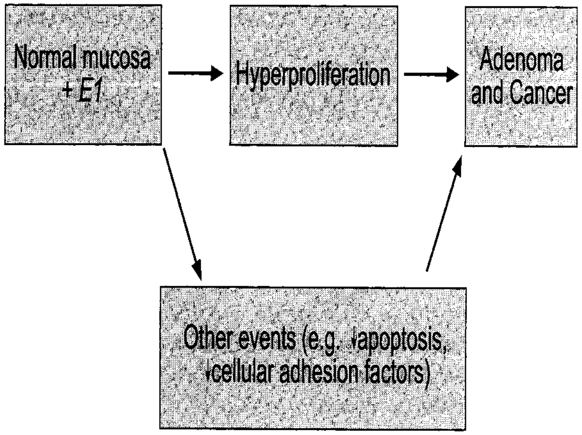

FIGURE 20.4. *Causal events potentially involved in the relationship between hyperproliferation and the neoplastic process in the colorectum.*

It is important to emphasize that the proliferation marker does not necessarily give the wrong answer about the agent's effect on cancer; the proliferation data might, in fact, be giving us the right answer. The problem is the uncertainty that flows from the existence of several alternative pathways to cancer.

20.8 Evaluating Potential Surrogate Endpoints

Given this uncertainty, how can we evaluate the validity of a potential surrogate marker? The answer is to integrate it into observational epidemiologic studies or clinical trials that have cancer (or a preneoplastic lesion, such as adenomatous polyps; see below) as an endpoint. This integration can elucidate the causal structure underlying the relationships among interventions (or exposures), potential surrogate endpoints, and cancer. In other words, the validation study should include data on T, S, and E for each individual and, if one wishes to demonstrate consistent ability to predict the magnitude of the effect of E on T from data on the effect of E on S (trial-level validity), there should be a series of such studies.

To determine whether the surrogate is valid for hypothesis testing, we need to investigate three questions:

1. Is the potential surrogate associated with cancer incidence (i.e., is S associated with T)?
2. Is the exposure or treatment associated with the potential surrogate (is E related to S)?
3. Does the potential surrogate endpoint "mediate" the relationship between exposure or treatment and cancer? That is, conditional on an individual's value of S, is there an absence of association between T and E, as in Figure 20.2a?

Standard epidemiologic measures such as relative risk and attributable proportion can be used in addressing these questions (Rothman and Greenland 1998).

20.8.1 Is the Surrogate Associated with Cancer?

As indicated above, for a marker to be a reasonable surrogate for a given cancer, it must be associated with that cancer. Ecologic studies can provide useful, if indirect, information on this connection. Studies are considered to be "ecologic," or aggregate, when individual-level information is not used; instead, an average marker value is obtained for a sample of individuals selected from specific populations (e.g., Seventh Day Adventists *versus* non-Adventists), which is then related to the overall risk of cancer in those populations. Several studies, for example, have compared mean proliferation indices in groups at varying risk of cancer (Lipkin *et al.* 1984). In such studies, however, one cannot be certain that those who are marker-positive are the ones with increased incidence of cancer.

This "ecologic" problem is obviated by moving to individual-level observational epidemiologic studies, whether case-control or cohort. Such studies give individual-level information on T, S, and E and they are important tools for examining the relationship between a putative surrogate and cancer. Blood estrogen levels have been shown in several studies to be directly associated with breast cancer, a relationship that had to be established before estrogens could be considered a surrogate for breast malignancy (Toniolo *et al.* 1995, Hankinson *et al.* 1998). Human papillomavirus (HPV) infection, a potential surrogate for cervical cancer, has been shown to be highly associated with risk of severe cervical neoplasia (Schiffman *et al.* 1993). Observational studies can also be incorporated into clinical trial design. For example, in the Polyp Prevention Trial (Schatzkin *et al.* 2000), a dietary intervention study with adenomatous polyp formation as the primary endpoint, investigators are currently examining the relationship between colorectal epithelial-cell proliferation measures and subsequent adenoma recurrence. The adenoma or CIN endpoints described here are only neoplastic cancer precursors; we have, for purposes of discussion, considered these as proxies for cancer, even though, as we discuss below, the validity of these precursor endpoints is not ironclad.

The attributable proportion (AP), an epidemiologic parameter that measures the extent to which T is determined by S, can be useful in determining the importance of alternative pathways and thereby evaluating the relationship between S and T. In the simple linear causal model of Figure 20.2a, the estimated AP for the surrogate is 1.0, excluding random error. When at least one pathway exists that is alternative to the pathway containing the surrogate, as in Figure 20.2b, then the AP for the surrogate is <1.0. A relatively high AP that was still less than 1.0, would suggest that the alternative ("$M2$") pathway plays a small role in tumorigenesis. An AP substantially lower than 1.0 for the surrogate implies that one or more alternative pathways is indeed operative, or that S is measured with a substantial degree of error (see Section 20.10).

20.8.2 Is E Associated with S?

Assuming that we are dealing with an intervention (exposure), E, that has an established relationship with T, for a potential surrogate marker to be valid, there must also be some relationship between E and the marker. Ecologic studies can provide indirect information on this question. For example, the mean colorectal epithelial cell proliferation index could be measured in populations with different average consumption of dietary fat. Individual-level studies, however, can provide more convincing evidence.

In a clinical trial, we need to see that the intervention changes the marker, which can be addressed in relatively small studies. Several studies, for example, have examined the effect of dietary change or supplementation on colorectal epithelial cell proliferation (Holt *et al.* 1998); others have investigated the effect of dietary fat modification (Prentice *et al.* 1990) or alcohol consumption (Reichman *et al.* 1993), both possible etiologic factors in breast cancer, on blood or urine estrogen levels. One illustrative case is that no relationship was found between calcium carbonate supplementation and epithelial cell proliferation measured one year later (Baron *et al.* 1995a), even though calcium did reduce overall adenoma recurrence (Baron *et al.* 1999). This suggests that proliferation measures are problematic surrogates for colorectal neoplasia/cancer in studies with calcium supplements as the main intervention/exposure.

We can also examine this question in case-control or cohort studies, in which we evaluate the association between an exposure and the potential surrogate. Schiffman *et al.* (1993), for example, in investigating the etiology of cervical cancer, showed a strong association between reproductive risk factors, particularly number of sexual partners, and HPV infection, a potential surrogate for cervical neoplasia. In a recent meta-analysis of cohort studies, body mass index was shown to be directly associated with blood estrogen levels (Endogeneous Hormones and Breast Cancer Collaborative Group 2003).

20.8.3 Does S Mediate the Link Between E and T?

Once we have determined (1) whether a potential surrogate is highly associated with cancer and (2) whether a surrogate is indeed linked to a given intervention or exposure, it is still necessary to determine whether (3) the effect of E on T is "mediated" by S in order to establish the validity of S for hypothesis testing. In statistical terms, mediation by S means that E and T are unrelated ("conditionally independent") once marker status is taken into account. One way to test for this condition is to stratify the data on levels of the surrogate marker and determine if there is an association between E and T within strata. If no such association is present, then there is evidence of mediation. An analogous approach is to include the surrogate marker S and the exposure E as independent variables in a multiple regression model that has T as the dependent variable. If the regression coefficient for E is 0, this constitutes evidence for mediation. The statistical aspects of mediation analysis are an area of current research (Freedman, Graubard, and Schatzkin 1992, Buyse and Molenberghs 1998).

We can obtain concrete data on mediation by integrating an assay for the surrogate into either clinical trials or observational epidemiologic studies, collecting information on both the intervention or exposure and the cancer (or severe neoplasia). As an example, investigators have used a case-control study to look at the extent to which HPV infection mediates the association between number of sexual partners and dysplasia (Schiffman and Schatzkin 1994). As Table 20.1 shows, the number of sexual partners was strongly and directly associated with cervical dysplasia risk. When the presence or absence of HPV infection was included as a covariate in a statistical regression model that related dysplasia to the number of sexual partners, the relative risk for number of sexual partners dropped dramatically. This suggests that most of the association between number of partners and cervical dysplasia is mediated through HPV infection (Franco 1991).

The same analytical strategy can be used to assess the extent of surrogate mediation in other study designs. For example, in the meta-analysis discussed above (Endogeneous Hormones and Breast Cancer Collaborative Group 2003), a direct association between BMI and breast cancer essentially disappeared after researchers adjusted for blood estrogen levels. A dietary modification or dietary supplement study of colorectal neoplasia, from which rectal biopsy specimens are obtained for mucosal proliferation assays, could provide information on the extent to which any observed diet/supplement effect is mediated by proliferation changes.

TABLE 20.1. *Number of sexual partners and the risk of cervical dysplasia.*

Odds ratio	Number of sexual partners				
	1	2	3–5	6–9	>10
Unadjusted	1.0	1.7	3.1*	4.7*	4.4*
Adjusted for HPV status	1.0	1.0	1.1	1.5	1.6

*: $p < 0.05$.

HPV, human papilloma virus.

As a general rule, the greater the intervention effect or exposure association, the fewer study participants are needed in a mediation analysis. For a number of reasons, the relative risks due to exposures in observational studies tend to be larger than the intervention effects observed in clinical trials. It follows that mediation analyses might be more likely to provide interpretable data in observational epidemiologic studies. Although complete mediation is necessary for a marker to be perfectly valid for hypothesis testing, it does not guarantee that the magnitude of the effects of E on S can be used to predict the magnitude of the effects of E on T reliably. Moreover, a demonstration that S mediates the effect of E on T for one exposure does not guarantee that it does so for another exposure. These points highlight the desirability of obtaining data on E, S, and T in several studies with possibly differing exposures.

20.9 Surrogates That Are Likely To Be Valid

Unlike putative surrogates such as epithelial cell proliferation or blood hormone levels, for which validity is problematic, considerable evidence supports the usefulness of a few "downstream" surrogate markers, that is, those close to cancer on the causal pathway.

Cervical cancer surrogates. Practically all cervical cancer requires prior persistent HPV infection. HPV persistence results in inactivation, by the E6 and E7 proteins of the HPV genome, of the *TP*53 and *RB* tumor suppressor genes, leading in turn to increasingly severe intraepithelial neoplasia and, eventually, cancer (zur Hausen 2000). At most only a very small proportion of cervical cancer can arise as a result of tumor suppressor inactivation occurring by mutation in the absence of HPV infection. Because most cervical cancer does occur through persistent HPV infection, an intervention that eliminates or reduces such infection would have a high likelihood of decreasing cervical cancer incidence.

Cervical intraepithelial neoplasia (CIN), especially CIN3, is also considered a strong surrogate for cancer and has been used as an endpoint in a number of epidemiologic studies. A very high percentage of CIN3 will progress to cancer in 20 years; only a very small fraction regresses. In fact, CIN3 is very close to being invasive cancer and is downstream from persistent HPV infection in the causal pathway leading to malignancy.

Adenomatous polyps for colorectal cancer. Another potential surrogate endpoint for which inferences to cancer are considered to be strong is the adenomatous polyp (adenoma). Colorectal adenomas are attractive candidates for cancer surrogacy in research studies because of their high recurrence rate: about 10% of persons having an adenoma removed will have a recurrence in the next year, an occurrence frequency nearly 2 orders of magnitude greater than the incidence of cancer. The underlying *biological* rationale for the use of adenoma endpoints in epidemiologic studies and clinical trials is the strong evidence for a relationship between this marker and colorectal cancer. This adenoma-carcinoma sequence is supported by studies demonstrating carcinomatous foci in adenomas and adenomatous foci within carcinomas, experiments showing the malignant transformation of adenoma cell lines, and studies identifying common mutations in adenomatous and carcinomatous tissue (Sugarbaker *et al.* 1985, Paraskeva *et al.* 1990, Fearon 1990). An intervention reducing the recurrence of adenomas in the large bowel would therefore probably decrease the incidence of colorectal cancer, thus making adenoma recurrence a reasonably valid surrogate marker.

Nevertheless, even the adenoma is not a perfectly reliable surrogate and some inferential difficulties remain with trials in which adenoma recurrence is used as a surrogate endpoint. Recurrent adenomas occur early in the tumorigenic sequence. The results of adenoma recurrence trials can be misleading if the intervention factor being tested operates later in the neoplastic process, for example from the growth of a small into a large adenoma or the transformation of a large adenoma to carcinoma. A (false) null result for recurrent adenomas can result if the intervention operates only in the later stages of neoplasia. A positive result, though, suggests that cancer would be reduced, because large adenomas and cancers derive from small adenomas.

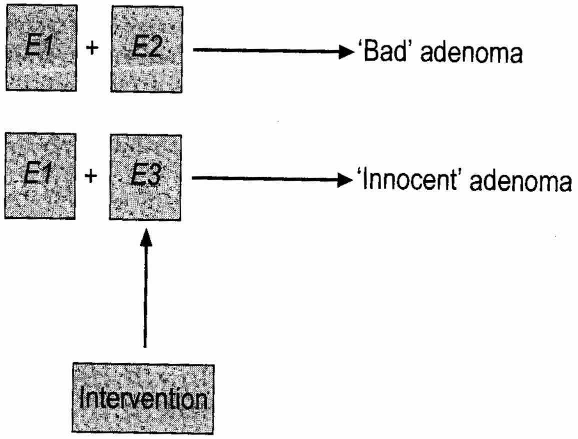

FIGURE 20.5. *Hypothetical setting.*

A second inferential difficulty with adenoma recurrence as a surrogate endpoint flows from the likely biological heterogeneity of adenomas. Only a relatively small proportion of adenomas go on to cancer. Suppose that one type, the "bad" adenoma that progresses to cancer, is caused by exposures $E1$ and $E2$, as in Figure 20.5. The second type, the "innocent" adenoma, is caused by the same exposure $E1$ but in concert with exposure $E3$. Imagine an intervention that works only on exposure $E3$. We could reduce the pool of innocent adenomas, thereby yielding a statistically significant reduction in adenoma formation in our trial, but in fact the incidence of bad adenomas and cancer would be unaffected. This could work the other way as well: we might see at most a small reduction in all adenomas (the bad ones being only a small proportion of all adenomas) even though the intervention truly decreases the formation of bad adenomas and, therefore, reduces the incidence of cancer.

20.10 Measurement Error

All biomarkers are measured with some error. Two important statistical issues need to be considered. First, a potential surrogate is useful (and ultimately valid) only if it can discriminate among study participants: those in the different treatment arms of a trial or the various exposure categories in an epidemiologic study. Discrimination is possible only if the surrogate values vary more between participants than they do within the same individual (for example, differences in marker values obtained from different tissue areas, measured at different time points, or read by multiple readers.) This can be measured by calculating a value known as the intraclass correlation coefficient (ICC), and this needs to be relatively large if the surrogate is to be useful (Fleiss 1986, pp. 1–5).

Intra-participant variability may be reduced, and the ICC thereby increased, by taking repeat samples, such as several biopsies from different areas or multiple blood samples over time. At a minimum, therefore, data are required on the potential surrogate marker's components of variance to establish the minimum number of marker samples needed for meaningful discrimination among study participants. In the absence of such data, it is not possible to ascertain whether null findings for a potential surrogate reflect a true lack of effect (or association) or simply the attenuating influence of random sources of intra-individual variation.

Reliability data have not been routinely collected in marker studies. Few studies have provided data on potential surrogate marker variability, particularly with respect to variability over time. A notable exception is recent investigations attempting to estimate the number of estradiol measurements necessary to discriminate among individuals (Hankinson *et al.* 1995). Studies measuring intra-individual variation in colorectal epithelial cell proliferation are under way (Lyles *et al.* 1994, McShane *et al.* 1998, Kulldorf *et al.* 2000). Quality-control studies designed to obtain data on the variability characteristics of potential surrogate markers are essential.

Second, even if the ICC is acceptable, measurement error will tend to attenuate findings from studies designed to answer each of the three questions posed above. The associations between intervention (exposure) and marker, and between marker and cancer, will be attenuated by errors in marker measurement (Franco 1991, Schiffman and Schatzkin 1994). Measurement error in S can also lead to an underestimate of the extent to which a correctly measured S would mediate the effect of E on T.

20.11 Conclusion

Because studies with surrogate cancer endpoints can be smaller, faster, and substantially less expensive than those with frank cancer outcomes, the use of surrogate endpoints is undeniably attractive. This attractiveness is likely to grow in coming years as the rapidly advancing discoveries in cell and molecular biology generate new therapies requiring testing and new markers that could plausibly serve as surrogates for cancer.

Surrogate endpoint studies can certainly yield useful information. They continue to play a legitimate role in Phase II clinical studies. In some areas of clinical therapeutics, surrogate endpoints like blood pressure, blood sugar level, or HIV viral load, are regarded as useful for Phase III studies. In other circumstances, the most that can be said is that surrogates *might* give the right answers about intervention effects on (or exposure associations with) cancer.

The problem is the uncertainty attached to conclusions based on surrogates. Except for those few surrogates that are both necessary for and relatively close developmentally to cancer, such as CIN3 and cervical cancer, the existence of plausible alternative pathways makes inferences to cancer from surrogates problematic. Merely being on the causal pathway to cancer does not in itself constitute surrogate validity; it is the totality of causal connections that is crucial. There is, unfortunately, a fairly extensive history of quite plausible surrogate markers giving the wrong answer about the effects of treatments for chronic disease (Fleming and DeMets 1996). There is no reason to believe that observational studies of cancer etiology based on cancer surrogates are immune to such inferential difficulties.

We should also consider the use of surrogate markers in the broader context of multiple disease endpoints, including treatment toxicity. A surrogate marker might give the "right" answer about cancer for a given intervention, but nevertheless give little or no information about important adverse events that greatly influence overall evaluation of the intervention. Suppose, for example, that we have a valid tissue or blood marker for breast cancer, one that gives us the right answer about a promising hormone-modulating intervention. That breast-cancer surrogate will tell us nothing about the potential of the intervention to increase the incidence of stroke. A potential stroke surrogate could be measured, but we are then faced with uncertainties about the reliability of this surrogate for stroke itself. This illustrates yet another difficulty arising from exclusive reliance on surrogate marker studies.

This chapter emphasizes the importance of conducting the investigations necessary to evaluate potential surrogates that include information on E, S, and T for study participants. Such studies are needed if we are to generalize from surrogate endpoint findings to cancer. There is, however, an implicit and perhaps unavoidable irony here: the large, long, expensive studies required to fully evaluate potential surrogates are precisely the studies that surrogates were designed to replace. Moreover, the exposure-dependence alluded to above complicates matters further: establishing validity for a given surrogate for one intervention/exposure does not necessarily translate into validity for another intervention/exposure. To assess validity for a variety of related interventions or exposures, the investigator needs a series of studies that provide individual-level data on T, S, and E.

The problems inherent in using surrogate endpoints need not be regarde as a cause for pessimism in cancer research. If anything, the limitations surrogacy remind us of the complexity of cancer causation and affirm th continued importance of large clinical trials and observational epidemi logic studies with explicit cancer endpoints. In the context of such a r search program, we may identify surrogates that can play a useful role i exploratory investigations and Phase II trials and, in some instances, i more definitive studies.